December 1946

BIOMETRICS

Vol. 2 NO. 6 BULLETIN

THE BIOMETRICS SECTION, AMERICAN STATISTICAL ASSOCIATION

WHICH REGRESSION?*

CHARLES P. WINSOR

The statistician has often to deal with the problem of fitting regressions when errors of measurement are present in one or both of the variables. Occasionally, some question arises as to which regression line is appropriate. Although this problem has been dealt with before, a good deal of confusion appears to exist, and occasional errors or misleading statements appear in textbooks. A very elementary presentation of a particularly simple case is offered here in the hope that it may be helpful to the biometrician.

The case which we consider is the following. A pair of variables u, v has a bivariate normal distribution in the general population, with variances σ_u^2, σ_v^2 and correlation ρ_{uv}. Our measurements of u and v are subject to error. We assume that these errors are independent, unbiased, and normally distributed. We actually record, then, individual measurements

$$x = u + \delta,\ y = v + \varepsilon,$$

where δ, ε are independent normal deviates with means zero and variances σ_δ^2, σ_ε^2. We are dealing, that is, with a case in which both errors of measurement and "organic variation" are present.

It is easy to see that in the general population the x, y measurements will be normally correlated; and the following relations can be shown to hold.

For the variances we have

$$\sigma_x^2 = \sigma_u^2 + \sigma_\delta^2\ ;\ \sigma_y^2 = \sigma_v^2 + \sigma_\varepsilon^2.$$

The correlation between x and y is

$$\rho_{xy} = \frac{\sigma_u \sigma_v}{\sigma_x \sigma_y}\,\rho_{uv}$$

* Paper No. 231 of the Department of Biostatistics, School of Hygiene and Public Health, The Johns Hopkins University.

The regression slopes in the general population are

$$\beta_{y \cdot x} = \frac{\sigma_u \sigma_v}{\sigma_x^2} \rho_{uv}\,; \quad \beta_{x \cdot y} = \frac{\sigma_u \sigma_v}{\sigma_y^2} \rho_{uv}.$$

Suppose now that we have a sample set of pairs of values of x and y. Which regression should we use, that of y on x, or that of x on y?

This question is meaningless as it stands. Before it can be answered, we must know (1) *how the* (x, y) *values were obtained* and (2) *what we are going to use the regression for.*

As to (1), there are two common situations.

(a) The pairs of (x, y) values were obtained as a random sample from the general population.

(b) The (x, y) values were obtained by selecting a set of values of one variable, say x, and subsequently measuring the corresponding values of y.

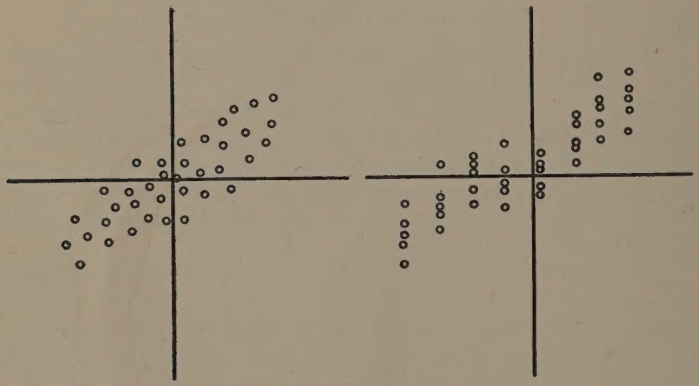

Figure 1. Left: Situation (a); x, y pairs randomly sampled. Right: Situation (b); x values arbitrarily chosen.

These two situations are indicated in the two panels of Figure 1.

As to (2), our proposed use of the regression line, there are four more or less usual possibilities.

(i) We may want a relation from which we can estimate in the future, the value of y, given a future measurement x.

(ii) We may want a relation from which to estimate x, given a future measurement y.

(iii) We may want a relation for estimating the true value v, given a future measurement x. (Or, u given y.)

(iv) We may wish to estimate the true relation between the true values u and v.

We shall consider first situation (a). Here, since our (x, y) pairs are a random sample from the population, we can estimate all the constants of the (x, y) distribution. With this information available, it is clear that we shall want the regression of y on x in problem (i), and that of x on y in problem (ii). This is, in fact, the classical bivariate normal regression problem, and is quite unaffected by the fact that errors of measurement are involved.

In problem (iii), it is easy to see that the regression of the true value v on the measurement x will be the same as that of the measurement y on x. This is so because y differs from v only by random and unbiased errors. Our best estimate of v will therefore be obtained from the regression of y on x.

For problem (iv), we need more information than is obtainable simply from a sample of (x, y) pairs. If we are trying to estimate the relation between the true values u and v, we need, in addition to the (x, y) pairs, estimates of the error variances σ_δ^2, σ_ϵ^2. (In some cases these are obtainable by duplicate measurements; but this is not always true.)

The physicists, in dealing with this problem, generally assume that their true values are perfectly correlated (functionally related), and that only errors of measurement are responsible for observational scatter. The computational techniques appropriate to this case are given fully by Deming (1943). The more general situation, where u and v may have any degree of correlation, has long been of concern to the psychologists. Spearman's "correction for attenuation" is one attempt to deal with it.

We now turn to situation (b), in which we selected a set of x values and measured the y's corresponding. Here we can obtain an estimate of the regression of y on x, and of the variance, $\sigma_{y \cdot x}^2$, of y around the regression line. We cannot, however, estimate the population values of $\bar{x}, \bar{y}, \sigma_x^2, \rho_{xy}$, nor can we estimate the population regression of x on y.

Entered as second-class matter, May 25, 1945, at the post office at Washington, D. C., under the Act of March 3, 1879. The Biometrics Bulletin is published six times a year—in February, April, June, August, October and December—by the American Statistical Association for its Biometrics Section. Editorial Office: 1603 K Street, N. W., Washington 6, D. C.

Membership dues in the American Statistical Association are $5.00 a year, of which $3.00 is for a year's subscription to the Quarterly Journal, fifty cents is for a year's subscription to the ASA Bulletin. Dues for Associate members of the Biometrics Section are $2.00 a year, of which $1.00 is for a year's subscription to the Biometrics Bulletin. Single copies of the Biometrics Bulletin are 60 cents each and annual subscriptions are $2.00. Subscriptions and applications for membership should be sent to the American Statistical Association. 1603 K Street, N. W., Washington 6, D. C.

The question is no longer "Which regression should we use?" but "What can we do with the single regression we have?"

Of the four problems previously considered, no difficulty arises in (i) or (iii), since in each of these the regression of y on x is required. We shall not consider problem (iv), which is obviously intractable under this situation. There remains for consideration problem (ii), to establish a relation for predicting x given y.

A careful discussion of this problem has been given by Eisenhart (1939) in terms of confidence intervals. We shall endeavor here to

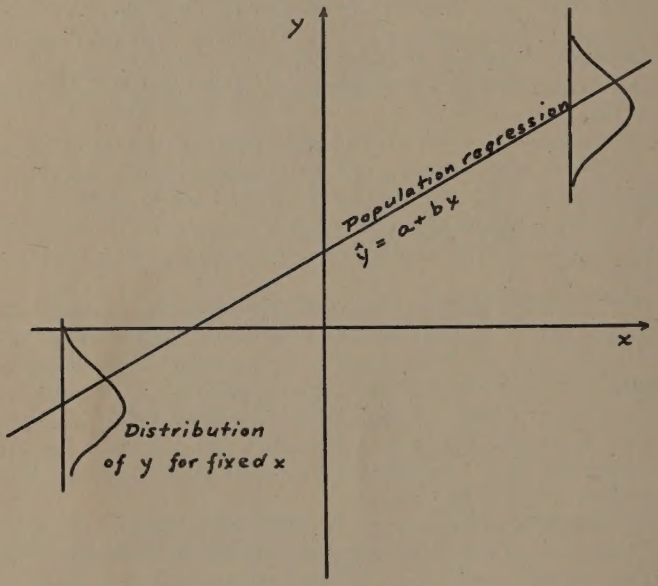

Figure 2. Population regression, y on x, and distribution of y about regression line.

concentrate on the fundamental problems of inference, avoiding the complications which arise out of the finite size of our samples. Let us assume, then, that we have been given an infinitely large sample, so that we know the exact regression of y on x, and the exact variance of y around the regression line, in the population. We do not know anything else. Suppose now a new (x, y) pair is taken from the

population and we are informed of its y value; what can we say about its x value?

At this point a diagram will be helpful.

We are given a regression line

$$\hat{y} = a + bx$$

(where we introduce the symbol y to distinguish the regression value from the individual sample values); and (since we are assuming normality) we know that for any assigned value of x, the values of y are

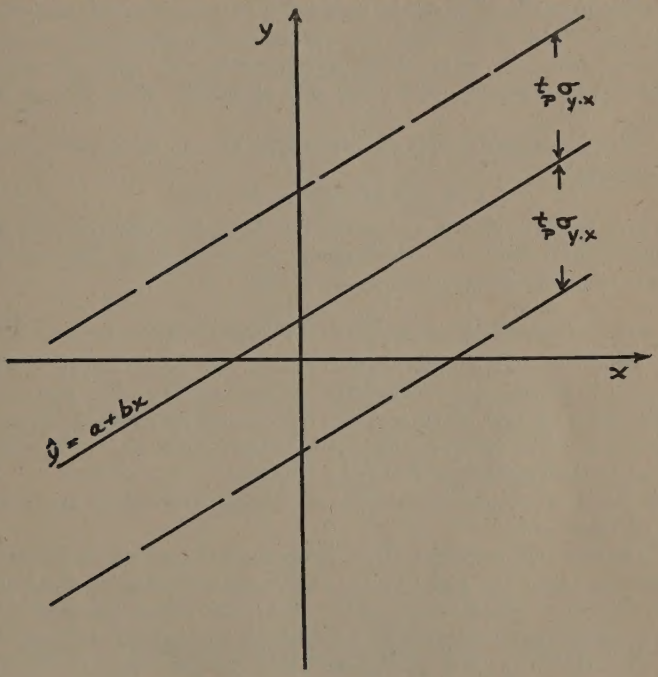

Figure 3. Confidence limits for y given x.

normally distributed with mean

$$\hat{y} = a + bx$$

and variance $\sigma_{y.x}^2$. We can, therefore, given any x, state the probability that the corresponding y shall fall within any assigned limits. In particular, for example, we can state that the probability that y lies in the interval

$$a + bx - 1.96\sigma_{y \cdot x} \text{ to } a + bx + 1.96\sigma_{y \cdot x}$$

is .95; and by proper choice of the coefficient of $\sigma_{y \cdot x}$, we can make corresponding statements for any other level of probability.

Again a diagram will be helpful.

We draw the regression line of y on x, and above and below it we draw lines at a vertical distance of $1.96\sigma_{y \cdot x}$. The probability that y falls between these two lines is .95 for every x value; and is therefore .95 for the totality of all possible y values. If, then, we consider all possible pairs of (x, y) values, we see that 95% of them will lie inside the strip which we have constructed; and that accordingly we can assert, given a value of y, that the corresponding x lies within the strip, and that this assertion will be true, over all possible cases, 95% of the time.

We can express all this algebraically. We can set up the double inequality

$$a + bx - t_P\sigma_{y \cdot x} < y < a + bx + t_P\sigma_{y \cdot x} \qquad \text{(A)}$$

with t_P properly chosen, and assert that the probability that this inequality is satisfied is P. Algebraically, we can rearrange this inequality to read as an inequality on x, and obtain

$$\frac{1}{b}(y - a - t_P\sigma_{y \cdot x}) < x < \frac{1}{b}(y - a + t_P\sigma_{y \cdot x}). \qquad \text{(B)}$$

Before we allow our algebra to run away with our judgment, it will be well to consider more closely the exact meanings of (A) and (B). With regard to (A), we observe that it is a statement about the random variable y, which involves a fixed but arbitrary value of x. The statement (A) has probability P of being true for any such arbitrary x. It has therefore the same probability P of being true for the aggregate of all values of x.

Consider now statement (B). If the random variable in this statement is y, then (B) and (A) are completely equivalent in meaning; but this is not what we really want. We should like to interpret (B) as a probability statement about an unknown x in terms of an observed y. But since (A) holds with probability P for the totality of all possible pairs (x, y), (B) will also hold with the same probability over all possible pairs. If, then, we make the assertion (B) about every (x, y) pair we draw, the assertion will be true with probability P—that is, a proportion P of our assertions will be true in the long run.

We cannot, however, say that in each particular case, or with respect to each particular y value, the assertion has probability P. In fact, it is easy to see that this is not true. For suppose that, knowing

the true regression, we set up our limits on x, in the form (B). And now suppose that in a long series of subsequent samples the value of x were (unknown to us) fixed and equal to x_0, say. The situation would then be that shown in Figure 4.

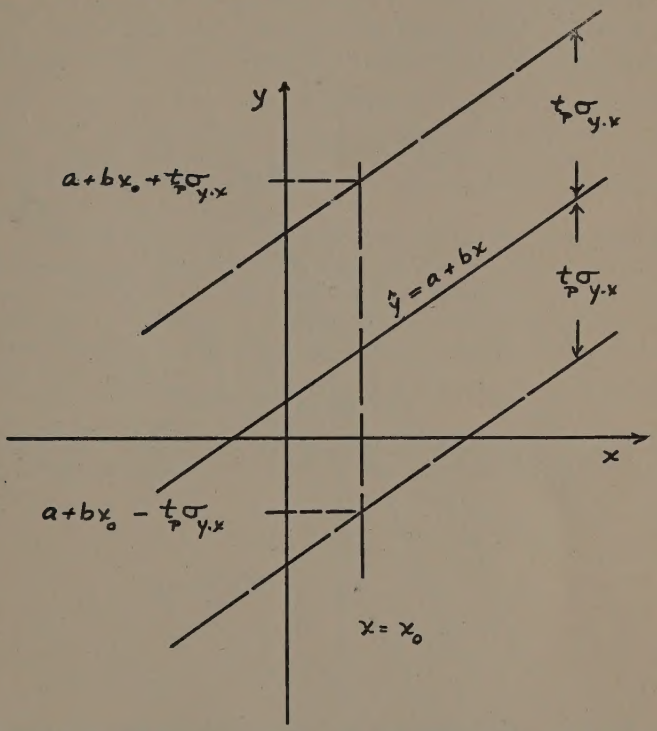

Figure 4. Illustrating meaning of inversion of confidence limits.

In this situation, whenever y falls outside the limits

$$a+bx \pm t_P\sigma_{y.x} \qquad \text{(C)}$$

our assertion (B) is false (it asserts that x lies within limits which do not in fact include the true value x_0) and whenever y falls inside the limits (C) our assertion (B) is true. Over the aggregate of all y values obtained, assertion (B) has probability P of being true, though for each particular y value it is either always true or always false.

The case we have just considered is an extreme one, chosen to illustrate the point. It corresponds to our initial assumptions about normal correlation if we make σ_x^2 zero. Figure 5 illustrates the situation for the general case of the bivariate surface. Here we have drawn both regression lines, that of y on x and that of x on y, and both sets of confidence limits. It will be noticed, first, that the two sets of confidence limits do not coincide, and second, that the direct limits are

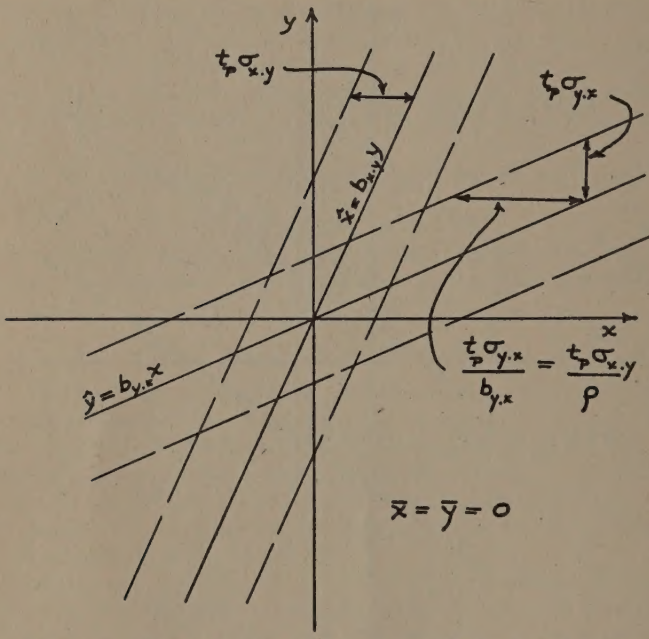

Figure 5. Regression lines and confidence limits, bivariate normal, x, y pairs randomly sampled.

narrower than the inverted limits. The advantages of using the direct regression when it is available are clear.

So far we have considered the case of an infinitely large sample. In practice our samples are of finite size, and accordingly our estimates of the regression line and of the variance around it are subject to sampling error. This results in more complicated algebra; in particular the confidence limits become hyperbolas, and in general we have

to face minor but somewhat troublesome complications of computation. Again, reference may be made to Eisenhart, where explicit formulae are given.

We may perhaps point out that this inversion of the regression line and the confidence limits is often the only available solution to the regression and estimation problem. For illustration we need only point to the case of biological assay. Here we are attempting to estimate the potency of an unknown in terms of a biological response, with the aid of a response curve based on known doses of a standard. The only possible regression line is that of response on dosage; even the notion of a population distribution of x values (potencies) becomes so vague as to be meaningless.

Our general principle, it appears, should be: if it is possible and meaningful, arrange the experiment so that the desired regression can be determined directly. That is, the variable from which prediction is to be made should be taken as the independent variable. In those numerous situations where this is not possible, use the inverted regression.

REFERENCES

Deming, W. E. *Statistical Adjustment of Data.* New York, John Wiley and Sons, 1943.
Eisenhart, C. The interpretation of certain regression methods and their use in biological and industrial research. *Annals of Mathematical Statistics,* 10 :162–186. 1939.

AN APPROXIMATE DISTRIBUTION OF ESTIMATES OF VARIANCE COMPONENTS

F. E. SATTERTHWAITE

General Electric Company, Ft. Wayne, Indiana

1. INTRODUCTION

In many problems, only simple mean square statistics are required to estimate whatever variances are involved. If the underlying populations are normal, these mean squares are distributed as is chi-square and may therefore be used in the standard chi-square, Student's t and Fisher's z tests. Frequently, however, the variances must be estimated by linear combinations of mean squares. Crump (1) has recently discussed a problem of this type, based on the following data:

ANALYSIS OF VARIANCE OF TOTAL EGG PRODUCTION OF 12 FEMALES *(D. melanogaster)* FROM 25 RACES IN 4 EXPERIMENTS

Source of Variation	*Degrees of Freedom*	*Mean Square*	*Average Value of the Mean Square*
Experiments	3	$MS_e = 46{,}659$	$\sigma_z^2 + 12\,\sigma_{er}^2 + 300\,\sigma_e^2$
Races	24	$MS_r = 3{,}243$	$\sigma_z^2 + 12\,\sigma_{er}^2 + 4\,\sigma_r^2$
E × R	72	$MS_{er} = 459$	$\sigma_z^2 + 12\,\sigma_{er}^2$
Within Subclasses	1,100	$MS_z = 231$	σ_z^2

The variance of the mean of the i th race is shown in his paper to be estimated by

$$(1) \qquad V_{.i.} = \frac{1}{e}\left(\hat{\sigma}_e^2 + \hat{\sigma}_{er}^2\right) + \frac{1}{en}\left(\sigma_z^2\right)$$

$$= \frac{1}{e}\left\{\frac{MS_e - MS_{er}}{300} + \frac{MS_{er} - MS_z}{12}\right\} + \frac{1}{en}\left(MS_z\right)$$

$$(2) \qquad = \frac{1}{e}\left\{\frac{MS_e}{300} + \frac{24\,MS_{er}}{300}\right\} + \left(\frac{1}{n} - \frac{1}{12}\right)MS_z\Bigg\}$$

where e is the number of experiments and n is the number of females in each experiment. Variance estimates such as (2) have been called *complex estimates* (2). Thus a complex estimate of variance is a linear function of *independent* mean squares.

It is stated in (1) that "increasing the number of females indefinitely still leaves us with

$$(3) \qquad V(\bar{x}_{.i.}) = \frac{MS_e + 24\,MS_{er} - 25\,MS_z}{300\,e} = \frac{173}{e}\,."$$

Conclusions are then reached without analysis of the sampling errors involved. Now the standard deviation of $V(\bar{x}_{.i.})$ is very large

$$(4)\qquad \{\hat{V}[\hat{V}(\bar{x}_{\cdot i\cdot})]\}^{\frac{1}{2}} = \frac{\sqrt{2}}{300\,e}\left[\frac{(MS_e)^2}{5} + \frac{(24\,MS_{er})^2}{74} + \frac{(25\,MS_s)^2}{1102}\right]^{\frac{1}{2}}$$
$$= 0.57\,\hat{V}(\bar{x}_{\cdot i\cdot})\,;$$

and further analysis leading to confidence limits for $V(\bar{x}_{\cdot i\cdot})$ should be helpful in choosing a course of action.

The writer has studied the distribution of complex estimates of variance in a paper (2) in *Psychometrika*. Since this paper may not be readily available to biometricians, the principal results are outlined below and a few applications are given.

2. The Distribution of Complex Estimates of Variance

The exact distribution of a complex estimate of variance is too involved for everyday use. It is therefore proposed to use, as an approximation to the exact distribution, a chi-square distribution in which the number of degrees of freedom is chosen so as to provide good agreement between the two. This is accomplished by arranging that the approximating chi-square have a variance equal to that of the exact distribution. If MS_1, MS_2, . . . are independent mean squares with r_1, r_2, . . . degrees of freedom and

$$(5)\qquad \hat{V}_s = a_1(MS_1) + a_2(MS_2) + \ldots$$

is a complex estimate of variance based on them, the number of degrees of freedom of the approximating chi-square is found to be given by

$$(6)\qquad r_s = \frac{[a_1E(MS_1) + a_2E(MS_2) + \ldots]^2}{\dfrac{[a_1E(MS_1)]^2}{r_1} + \dfrac{[a_2E(MS_2)]^2}{r_2} + \ldots}$$

where $E(\ \ )$ denotes mean or expected values.

In practice, the expected values of the independent mean squares will not be known. The observed values will usually be substituted in (6), giving, as an estimate of r_s,

$$(7)\qquad \hat{r}_s = \frac{[a_1(MS_1) + a_2(MS_2) + \ldots]^2}{\dfrac{[a_1(MS_1)]^2}{r_1} + \dfrac{[a_2(MS_2)]^2}{r_2} + \ldots}.$$

An approximation of this type for a slightly simpler problem was first suggested by H. Fairfield Smith (3). In his problem, there were only two mean squares and $a_1 = a_2 = 1$. This approximation does not support the use of $r+2$ in place of r as a correction for bias [(1) formula 3].

The writer has checked the accuracy of the suggested approximations by calculating the exact distribution for a number of special cases. Typical results are as follows:

r_1	r_2	$\frac{E(MS_1)}{E(MS_2)}$	r_s	$\chi^2(95\%)$		$\chi^2(99.9\%)$	
				exact	approx.	exact	approx.
4	2	4	100/33	7.9	8.0	16.2	17.3
8	4	1	32/ 3	19.4	19.5	30.5	31.0
6	4	2	54/ 7	15.1	15.3	26.0	27.2
20	4	2	180/21	16.2	17.0	27.7	29.0
4	2	1	16/ 3	11.5	11.7	21.3	22.3

The above discrepancies between the exact and the approximate chi-squares, even for the extreme 99.9 percent case, are very small compared with their sampling errors. Thus it appears that the approximation may be used with confidence. Furthermore, we know from general reasoning that if r_s is large, both the approximate and the exact distributions approach the same normal distribution; if r_s is small, the sampling errors in the chi-squares are large and refinement is superfluous.

Some care must be taken in the cases where one or more of the a's in (5) are negative. If it is possible for $\hat{V}_s$ to be negative with a fairly large probability, the approximate distribution will become rather poor since it can not allow negative estimates. However, here again the sampling errors in $\hat{V}_s$ will be quite large compared with its expected value so that only the sketchiest of conclusions can be drawn in any case.

3. Further Analysis of Crump's Example

The distribution of Crump's estimate of the residual variance of the race means,

$$\hat{V}(\bar{x}_{\cdot i\cdot}) = \frac{(MS_e) + 24(MS_{er}) - 25(MS_s)}{300e} \tag{3}$$

can now be approximated. Thus

$$\hat{V}(x_{\cdot i\cdot}) = \frac{1}{e}\left[\frac{46{,}659}{300} + \frac{(24)(459)}{300} - \frac{(25)(231)}{300}\right] \tag{8}$$

$$= \frac{1}{e}[155 + 37 - 19] = \frac{173}{e}.$$

From (7) we have

$$\hat{r}_s = \frac{[155 + 37 - 19]^2}{\frac{(155)^2}{3} + \frac{(37)^2}{72} + \frac{(19)^2}{1{,}100}} \tag{9}$$

$$= \frac{29{,}929}{8{,}008 + 19 + 1} = 3.7$$

From chi-square tables interpolated for 3.7 d.f. at the 5 percent and 95 percent points we find that, with a high degree of probability,

(10) $$0.60 < \frac{3.7\,\hat{V}}{V} < 9.0$$

or

(11) $$\frac{(3.7)(173)}{(9.0)e} < V < \frac{(3.7)(173)}{(0.60)e}$$

$$\frac{71}{e} < V < \frac{1{,}067}{e}.$$

Thus if it were necessary to reduce V to 9 and if time were important so that a second series of experiments could not be made, we should run

(12) $$e = \frac{1{,}067}{9} = 119$$

experiments in the first series for confidence that V would be properly reduced. On the other hand, if the experiments were expensive and time not important, we might run

(13) $$e = \frac{(3.7)(173)}{(5.6)(9)} = 13$$

experiments and then get a more accurate estimate of V to determine how many additional experiments should be run (5.6 obtained from the 20 percent point for chi-square, 3.7 d.f.).

4. Difference of Means

The usual estimate of variance used in Student t tests for the difference of two means is

(14) $$V_t = \left[\frac{r_1(MS_1) + r_2(MS_2)}{r_1 + r_2}\right]\left[\frac{1}{r_1+1} + \frac{1}{r_2+1}\right]$$

with

(15) $$r_t = r_1 + r_2$$

degrees of freedom. This assumes that both populations have the same variance. Seldom do we have positive evidence that this is so and often we have evidence that the variances are different. For example, $F = (MS_1)/(MS_2)$ may be significant. Note that a non-significant F is not evidence that the variances are equal, especially if one of the MS's has a small number of degrees of freedom.

The assumption of equal variances can be avoided by use of a complex estimate of variance,

(16) $$\hat{V}_s = \frac{MS_1}{r_1+1} + \frac{MS_2}{r_2+1}$$

with

$$(17) \qquad \hat{r}_s = \frac{\{[MS_1/r_1+1)] + [MS_2/(r_2+1)]\}^2}{\frac{[MS_1/(r_1+1)]^2}{r_1} + \frac{[MS_2/(r_2+1)]^2}{r_2}}$$

degrees of freedom.

For example, consider the numerical case:

$$MS_1 = 100, r_1 = 99,$$
$$MS_2 = 90, r_2 = 9.$$

By the standard analysis one would obtain

$$(18) \qquad \hat{V}_t = \left[\frac{(99)(100)+(9)(90)}{108}\right]\left[\frac{1}{100}+\frac{1}{10}\right] = 10.9$$
$$r_t = 99+9 = 108$$

The complex estimate gives

$$(19) \qquad \hat{r}_s = \frac{100}{100}+\frac{90}{10} = 10.0,$$
$$\hat{r}_s = \frac{(1+9)^2}{\frac{1^2}{99}+\frac{9^2}{9}} = 11.1.$$

One will sometimes reach different conclusions with 108 degrees of freedom from those he will reach with 11 degrees of freedom.

If from general reasoning or other *a priori* considerations it is believed that both MS_1 and MS_2 are independent estimates of the same variance, then the use of 108 degrees of freedom is justified. On the other hand, if the given data are the entire admissible knowledge, then the use of more than 11 degrees of freedom is not valid.

5. Conclusion

In many practical problems the most efficient estimate of variance available is a linear function of two or more independent mean-squares. Usually the exact distribution of such estimates is too complicated for practical use. A satisfactory approximation can be based on the chi-square distribution with the number of degrees of freedom determined by (7).

Many problems, such as the difference of means, can be more conservatively analyzed by use of complex estimates of variance. Assumptions regarding homogeneity of variance can then be avoided.

REFERENCES

(1) Crump, S. Lee. The estimation of variance components in the analysis of variance. *Biometrics Bulletin* 2:1:7–11. February 1946.

(2) Satterthwaite, Franklin E. Synthesis of variance. *Psychometrika* 6:309–316. October 1941.

(3) Smith, H. Fairfield. The problem of comparing the results of two experiments with unequal errors. *Journal of the Council of Scientific and Industrial Research* 9:211–212. August 1936.

PRELIMINARY REPORT ON THE RECTANGULAR LATTICES

Boyd Harshbarger

Virginia Agricultural Experiment Station and
Virginia Polytechnic Institute

The theory of the experimental arrangement which is now called the lattice had its beginning with Yates in 1936. It has been further developed by Yates and Cochran in numerous papers beginning in 1939 and extending through 1943. The lattice design requires that the number of varieties be an exact square. Within each replication, the varieties are laid out in incomplete blocks so that each row or incomplete block contains the same number of varieties. The grouping of varieties into blocks in the various replications is made according to a set of patterns which are so arranged that, in them, no pair of varieties occurs together within a block more than once. Each pattern may be employed one or more times (usually two). These patterns are referred to hereafter as Group X, Group Y, etc. This symmetry makes it possible to adjust the varietal total or average yields by simple calculations for variations in the fertility among the incomplete blocks.

To avoid the restriction that the number of varieties must be a perfect square, Yates introduced a design in which the blocks vary in size from one replication to another. This he called the pseudo-factorial with unequal groups of sets. However, no attempt was made to use inter-block information, and the design proposed by him does not conveniently lend itself to such an analysis.

This paper presents a few of the preliminary results on incomplete block designs in which the number of varieties is the product of two consecutive integers. The arrangement differs from Yates' non-square design since the blocks are all the same size and the variety means are adjusted for both the inter- and intra-block information. It has a closer resemblance to the ordinary lattice design than does the pseudo-factorial with unequal groups of sets. The name Rectangular Lattice is proposed for this design since the word lattice carries no implication of squareness.

The general theory of Rectangular Lattices is to be published in a Virginia Agricultural Experiment Station Bulletin which will also carry numerical examples of simple and triple rectangular lattices. Copies of this bulletin will be available on request.

As in the square lattice, the varieties are arranged in two groups, X and Y, each of which is replicated as shown below. For simplicity of illustration, a 3×4 rectangular lattice is used. The numbers designate varieties.

TABLE I

GROUP X

Blocks				*Blocks*			
(1)	1	2	3	(1)	1	2	3
(2)	4	5	6	(2)	4	5	6
(3)	7	8	9	(3)	7	8	9
(4)	10	11	12	(4)	10	11	12

GROUP Y

Blocks				*Blocks*			
(1)	1	4	7	(1)	1	4	7
(2)	5	8	10	(2)	5	8	10
(3)	2	9	11	(3)	2	9	11
(4)	3	6	12	(4)	3	6	12

In practice the varieties are randomized within blocks and the blocks are randomized within the replicates.

For purposes of enumeration and computation the rectangular lattice may be thought of as a square lattice with k varieties missing in such a way that the missing varieties occur once in each row and once in each block. With this arrangement the groups for a $k(k-1)$ rectangular lattice (together with more convenient subscripts) are shown in Table II.

TABLE II

GROUP X

Blocks						
(1)	v_{11}	0	v_{13}	$\cdots$ $v_{1,k-1}$	$v_{1,k}$	B_{x1}
(2)	v_{21}	v_{22}	0	$\cdots$ $v_{2,k-1}$	$v_{2,k}$	B_{x2}
(3)	v_{31}	v_{32}	v_{33}	$\cdots$ $v_{3,k-1}$	$v_{3,k}$	B_{x3}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$(k-1)$	$v_{k-1,1}$	$v_{k-1,2}$	$v_{k-1,3}$	$\cdots$ $v_{k-1,k-1}$	0	$B_{x,k-1}$
(k)	0	$v_{k,2}$	$v_{k,3}$	$\cdots$ $v_{k,k-1}$	$v_{k,k}$	$B_{x,k}$
	T_{x1}	T_{x2}	T_{x3}	$\cdots$ $T_{x,k-1}$	$T_{x,k}$	

Blocks

(1)	v_{11}	v_{21}	v_{31}	$\cdots v_{k-1,1}$	0	B_{y1}
(2)	0	v_{22}	v_{32}	$\cdots v_{k-1,2}$	$v_{k,2}$	B_{y2}
(3)	v_{12}	0	v_{23}	$\cdots v_{k-1,3}$	$v_{k,3}$	B_{y3}
	⋮	⋮	⋮	⋮	⋮	
$(k-1)$	$v_{1,k-1}$	$v_{2,k-1}$	$v_{3,k-1}$	$\cdots v_{k-1,k-1}$	$v_{k,k-1}$	$B_{y,k-1}$
(k)	$v_{1,k}$	$v_{2,k}$	$v_{3,k}$	$\cdots 0$	$v_{k,k}$	$B_{y,k}$
	T_{y1}	T_{y2}	T_{y3}	$\cdots T_{y,k-1}$	$T_{y,k}$	

Table II serves a double purpose. The body of the table gives the pattern of the arrangement in the blocks of the $k(k-1)$ varieties, the v's being simply symbols for varieties. However, if each v is regarded as the total of the two observations on a variety over the two replicates of a group, then this table is one which is made up in the course of the analysis. The marginal totals, the B's and T's, are formed on this basis. Thus, B_{xi} is the sum of the yields of the varieties which occur in the i^{th} blocks of the two replications in Group X. T_{xi} is the sum of the yields, in Group X, of the varieties listed in block i of Group Y, etc.

By a rather tedious mathematical process, which is given in detail in the Experiment Station Bulletin, the following formulas and equations are evolved. R_e is the sum for replicate e, A_{hi} is the difference between blocks within Group h of block i, V_i is the sum of variety i from the four replicates. G is the grand total, and y_{eij} is the individual observation.

The weights are calculated by two simple formulas

$$\frac{1}{W} = \frac{5(3k^2-7k-1)^2Q - 4(k-1)^2(N-2P)}{5(3k^2-7k-1)^2+4(k-1)^2}$$

and

$$\frac{1}{W'} = \frac{1}{5}\left(4N + 2P - \frac{1}{W}\right)$$

where N represents the mean square of component (a)

P " " " " " " (b)

Q " " " " " " the error.

TABLE III

ANALYSIS OF VARIANCE TABLE

Sources of variation	d/f	Sum of squares
Replicate	3	$\sum_{e=1}^{4} \frac{R_e^2}{k(k-1)} - \frac{G^2}{4k(k-1)}$
Component (a)	$2(k-1)$	$\sum_{h=x}^{y} \sum_{i=1}^{k} \frac{A_{hi}^2}{2(k-1)} - \frac{(R_1-R_2)^2+(R_3-R_4)^2}{2k(k-1)}$
Component (b)	$2(k-1)$	$\frac{1}{4k(k-2)} \left\{ \sum_{i=1}^{k} (k-1)[(B_{xi}-T_{yi})^2+(B_{yi}-T_{xi})^2] \right.$ $-2\sum_{i=1}^{k}(B_{xi}-T_{yi})[B_{y(i+1)}-T_{x(i+1)}]$ $\left. -2[(R_1+R_2)-(R_3+R_4)]^2 \right\}$
Varieties	$k(k-1)-1$	$\frac{\Sigma V_i^2}{4} - \frac{G^2}{4k(k-1)}$
Error (residuals)	$3k^2-7k+1$	by subtraction
Total	$4k(k-1)-1$	$Sy_{eij}^2 - \frac{G^2}{4k(k-1)}$

where the subscript $k+1$ is to be replaced by unity when it occurs.

The variety means are adjusted by using both the inter- and intra-block weights. This is accomplished by calculating certain constants and subtracting them from the variety averages. If the averages are arranged in the order of the x group, then the constants to be subtracted from row i and column i are as follows:

$$c_{xi} = \frac{W-W'}{4[(k-1)^2(W+W')^2-(W-W')^2]} \times \{(k-1)(W+W')(B_{xi}-T_{yi}) - (W-W')(B_{y(i+1)}-T_{x(i+1)})\}$$

where the subscript $k+1$ is to be taken as unity, and

$$c_{yi} = \frac{W-W'}{4[(k-1)^2(W+W')^2-(W-W')^2]} \times \{(k-1)(W+W')(B_{yi}-T_{xi}) - (W-W')(B_{x(i-1)}-T_{y(i-1)})\}$$

where the subscript zero is to be taken as k.

The standard error of the adjusted varietal means. The standard error of the difference between the means of two varieties occurring together in the same block is

$$\frac{1}{2W}\left[1+\frac{(k-1)(W-W')(W+W')}{[kW+(k-2)W'][(k-2)W+kW']}\right]$$

For two varieties not occurring together in the same block the standard error of the difference between the means is

$$\frac{1}{2W}\left[1+\frac{(W-W')[(2k-1)W+(2k-3)W']}{[kW+(k-2)W'][(k-2)W+kW')]}\right]$$

The average standard error of all varietal comparisons is

$$\frac{1}{2W}\left\{1+\frac{W-W'}{k^2-k-1}\right.$$
$$\left.\times\left[\frac{2(k-1)(k-2)(W+W')+(k^2-3k+3)[(2k-1)W+(2k-3)W']}{[kW+(k-2)W'][(k-2)W+kW']}\right]\right\}$$

The efficiency relative to randomized complete block design for rectangular and square lattices is given in the table below for different values of W/W'. The upper figures give the percentage efficiency for rectangular lattices with the number of varieties being consecutive integers, and the lower figures give the same for square lattices. The size of the lattice and whether it is a square or rectangular lattice is given in the left hand column of the table.

TABLE IV

PERCENTAGE EFFICIENCIES OF DESIGNS OF RECTANGULAR LATTICES AND SQUARE LATTICES RELATIVE TO RANDOMIZED COMPLETE BLOCKS

	W/W'						
	1	2	3	4	6	8	10
5×4	100	106	116	128	154	181	208
5×5	100	105	114	125	148	172	196
6×5	100	105	114	124	147	171	195
6×6	100	104	112	122	142	164	185
7×6	100	104	112	121	142	163	184
7×7	100	104	111	120	138	157	176
8×7	100	104	111	119	137	156	176
8×8	100	103	110	118	134	152	169
9×8	100	103	110	117	134	151	169
9×9	100	103	109	116	131	147	163
10×9	100	103	109	116	131	147	163
10×10	100	103	108	115	129	143	158

(39)

QUERY: I have been informed that there is a formula,

$$N = s^2\left(\frac{t}{m-\bar{x}}\right)^2,$$

to find the number of samples necessary to take within a given percentage of error.

It happens that I am interested in a problem of this kind and I was using the formula

$$N = 2t^2s^2/d^2,$$

(Soil Sci., Vol. 58:275–288, 1944) which gives results twice as high as the former. Will not the latter formula give a better approach to the problem?

ANSWER: The two formulas, giving answers to two different questions, cannot be used alternatively. Let us consider them separately.

First. A sample of n observations provides a value of s which indicates unsatisfactory reliability. How large a sample is required, drawn from the same population, to estimate the population mean with fiducial limits no greater than $x \pm d$ where $m - \bar{x} = d$? The first formula is a rough approximation to this sample size.

Second. Two samples are drawn, each consisting of n observations. The pooled standard deviation, s, and the difference between the means, d, do not indicate significance. How large should these samples be to show significance, assuming the population difference to be not less than d? The second formula, quoted by Cline, gives a rough approximation to N.

In each formula the assumption is made that the odds are about 50–50 that the prospective sample will correctly answer the question posed. If you wish to be more confident of the outcome, larger samples are required (see query next following).

Incidentally, if you got one result twice the other, I suspect an incorrect substitution of t in the formulas. In the first, t has degrees of freedom, $n-1$, while in the second, $d.f. = 2(n-1)$.

GEORGE W. SNEDECOR

(40)

QUERY: I have a sample of 35 observations with mean, 24, and standard deviation, 6. I calculated the half confidence interval, $st_{.05}/\sqrt{n} = 2.061$, so that I cannot be reasonably certain of being within less than 8.6% of the population mean. What sample size do I need

to be assured that my sample mean will be within 1.2 points of the population mean, 1.2 being 5% of the present sample mean?

ANSWER: The formula for the prospective sample size is

$$\sqrt{N} = (s/d)t\sqrt{F}$$

where d is the desired half confidence interval, 1.2, t is at any specified level (0.05, say) with $N-1$ degrees of freedom, and F has the degrees of freedom, $n_1 = N-1$ and $n_2 = 35-1 = 34$. F may be set at the tabulated point which gives you satisfactory assurance that your proposed sampling will be successful.

Since N, t and F all pertain to the sample whose size is to be determined, solution of the equation must be by successive approximation. From the first sample, together with your specification of the half interval, $s/d = 6/1.2 = 5$. For any contemplated sample above 30, t can be set at 2. Substituting,

$$\sqrt{N} = (5)(2)\sqrt{F}, \text{ or } N = 100F$$

With this approximate relation, follow the line for $n_2 = 34$ in the F table until you find $N = 100F$. If you are using Snedecor's table and $P = 0.05$, you will observe that, at $n_1 = 100$, $N = 101$ is less than $100F = 164$; while at $n_2 = 200$, $N = 201$ is greater than $100F = 161$. Linear interpolation places N at about 163. This means that if you draw a sample of 163 from the same population as before, the probability is 0.95 that the new half interval will be 1.2 or less. As to whether this value is more or less than 5% of the new sample mean is not specified.

In rare instances when one needs improvement of the foregoing approximation, two refinements can be introduced. (i) Substitute the more accurate value of t for $d.f. = 162$; that is, $t_{.05} = 1.975$. (ii) Instead of linear interpolation, use the method described by Fisher in section 41 of his "Statistical Methods;" that is, plot F against the reciprocal of n_1, then interpolate linearly.

You may not demand such assurance as 19 in 20 that the new sample will turn up $d = 1.2$ or less. If you are satisfied with chances of 4 in 5, you can use the 20% points of F (variance ratio) in the Fisher and Yates table. Other probabilities are tabulated by Merrington and Thompson in Biometrika, Vol. 33, pp. 73–88 (1943).

The approximation described for the preceding query is based on the value, $F = 1$, the assumption being that this is the 50% value of F. The closeness of the approximation can be assessed by examination of the 50% points in Merrington and Thompson's table. $F_{.5} = 1$ only if $n_1 = n_2$, but the approximation is generally satisfactory if both n_1 and

n_2 are greater than 20. Illustration: approximate N from the data given above but with F at its 50% point. A double interpolation is required, as in Fisher's example, the result being $N = 100$. The interpretation is that a second sample of 100 observations is as likely as not to yield $d = 1.2$ or less.

This should be compared with the solution given by the first formula in the foregoing query. Since $t = 2.032$ for $d.f. = 34$, we have

$$N = (2.032)^2(6)^2/(1.2)^2 = 103$$

The excellence of this approximation is due to: (i) the small discrepancy between the original value of t for 34 $d.f.$, 2.032, and the correct value for the new sample, $t = 1.984$, $d.f. = 99$; and (ii) the close approach of the 50% F to 1, its value in this example being 1.013.

A. M. Mood and George W. Snedecor

(41)

QUERY: As a "practical statistician" I sometimes wonder about the justification of utilizing certain refinements in practical every-day statistics. My reaction is that a practical statistician, to be competent, *should* know statistical method; but, to be practical, he should also realize that crude methods may often adequately serve immediate needs. Too, when errors are so very easy to enter, it sometimes seems to me that precise methods are of questionable practical utility. Broadly, and specifically, it seems to me that in dealing with any but relatively small samples in "practical" fields, utilizations of such devices as degrees of freedom and other similar adjustments are not generally warranted. Perhaps, I might express myself more clearly if I were to say that it seems to me that such refinements have definite limitations and should not be used indiscriminately. I'd appreciate your brief comments.

ANSWER: I can agree to most of your comments—the experienced statistician can often anticipate the outcome of an investigation quite accurately and thus avoid tedious computation. Or, he may be confident that a simple method will extract all the information necessary for practical purposes, thus avoiding elaborate processes. A colleague of mine refers to this as "zoot" statistics, a kind used freely by competent investigators.

Some research men tolerate rough treatment of their data while a few insist upon it; but my experience has been that most of them are satisfied with nothing less than complete extraction of available information. Can one blame them? After they have labored long over

exacting techniques for incorporating information in their data, there is every reason for using the most efficient statistical methods for eliciting the last ounce of it. Usually the statistical work, even the most complicated, is trivial compared to the time, energy, and money that have gone into the experimental work.

Doesn't it all boil down to this: how much is the information worth, in dollars and cents and human energy? Certainly one should discourage the expenditure of money or effort if there is clearly little information in the data, or if the results can obviously have little value. But data which are packed with expensive and useful information warrant the utmost refinement in statistical methods of extraction.

GEORGE W. SNEDECOR

(42)

QUERY: In certain types of investigation it is possible to make rather accurate observational classifications, but satisfactory physical measurements have not been developed. Numerical values may be assigned to these classes, although because of lack of physical-measurement standards such values may bear no exact proportional relation to the characters observed. The distribution of the values may be approximately normal.

Is it permissible to treat such data by analysis of variance (1) when the distribution is normal, and (2) when the distribution deviates from normal?

ANSWER: Several distinctions should be made before any answer has meaning.

First. Analysis of variance is merely an arithmetical process of allocating variance to observed classes and, as such, is applicable to any numerical data. Contrariwise, probability statements, such as those involved in tests of significance, are affected by the distribution of the measured variate. I assume that the query is about the accuracy of the probability turned up in an F test of significance.

Second. Observational classifications are opinions of the observer based on physical characteristics of the material being judged. Presumably querist is asking whether the test of significance will lead to correct conclusions about the *material*, irrespective of the observer. This involves the appropriate design of the sampling: it must include independent observations by at least two persons who have the ability and the training required to align their opinions with facts. An allowance for variation among opinions must be made available by the analysis of variance and must be included in the estimate of error.

Third. The assignment of numbers to the observational classes is understood to be governed by the physical characteristics of the observed material; otherwise, conclusions based on such numbers may be unrelated to the phenomenon being investigated. Any statistical procedures, such as averages, analysis of variance and related tests of significance, are futile if the numbers used do not measure the variate being studied. In certain large fields of investigation, notably in mental measurements, there are reasons to believe that the numbers should be assigned so that the resulting distributions are normal; but this, while statistically convenient, is clearly not universal.

With these distinctions in view, my answer to the first question is "yes." The second question introduces difficulties in statistics, some of which were discussed in Query No. 32, Vol. 2, No. 4, p. 73–74 of this Bulletin. If anormality is extreme, the probability associated with the test of significance may be seriously affected: the advice of a mathematical statistican should be sought.

GEORGE W. SNEDECOR

Eight new members have been added to the staff of the Statistical Laboratory, Iowa State College. They are: GEORGE W. BROWN from the R.C.A. laboratories; LEONID HURWICZ recently research associate Cowles Commission, Chicago; ALEXANDER M. MOOD, who was research associate at Princeton; HERMAN O. DRABENSTOTT, CLIFFORD J. MALONEY and NORMAN V. STRAND who have had former associations with this laboratory; JAMES G. DARROCK formerly agricultural scientist, Dominion Laboratory of Cereal Breeding, Winnipeg; and GARNET E. McCREARY who received his M.A. from Queen's in 1946. The new bi-monthly publication, "Statlab Review," by the Iowa State College Statistical Laboratory, has as its editor, JOSEPH C. DODSON. This paper gives a brief report of the two "Allied Missions to Observe the Greek Elections." . . . After returning from two years of duty as an Operations Analyst with the Army Air Forces in the European Theater, A. E. BRANDT returned to the Soil Conservation Service as Research Specialist in charge of experimental design and analysis of data. On October 1, he reported to the Naval Ordnance Laboratory as Statistical Consultant on the staff of the Technical Director of the laboratory. MRS. BRANDT was recently thrown from one of their horses and broke her hip. Our sympathy and may the recovery be speedy. . . . T. A. BANCROFT is now on the staff of the Department of Mathematics, The University of Georgia, Athens. He writes, "I am organizing courses in statistics here and starting consulting work for various research workers. The University of Georgia Science Club held a symposium on 'The contribution of statistical methods to research' on Tuesday, November 26. In the evening at eight o'clock, W. G. COCHRAN delivered the principal address. His address was preceded by a dinner in his honor given by members of the Science Club. In the afternoon at four o'clock a number of short addresses were given on the necessity for statistical methods in research in the various fields of applied science. These talks were given by B. O. WILLIAMS, A. S. EDWARDS, W. T. HICKS, EDWIN JAMES and CHARLES C. WILSON. I spoke on 'Statistics as a mathematical subject.'" . . . WARREN H. LEONARD has returned to Colorado A & M College, Fort Collins, after an absence of four years in the army. . . . CHARLES P. WINSOR, Department of Biostatistics, Johns Hopkins University, who resides at 615 North Wolfe Street, Baltimore, Maryland, is editor of Human Biology. He would particularly like to get papers which

are of interest from the point of view of quantitative methodology. The editor of this Bulletin extends condolences. We need articles too! . . . The Columbia Broadcasting System has started a series of radio discussions on "You and Alcohol" with initial discussion by E. M. JELLINEK, biometrician and director of the Section on Alcohol Studies of the Laboratory of Applied Physiology, Yale. . . . ALFRED SAUVY, Director, National Institute of Demographic Studies, 20, Rue de la Baume, Paris, with his wife and PAUL E. VINCENT visited the Institute of Statistics at Raleigh November 5. They were interested in seeing the cotton-growing experiments and in hearing and watching a tobacco auction. . . .

At the Annual Meeting of the Animal Vitamin Research Council held in Washington October 17, 1946, it was voted to change the name of the organization to the Animal Nutrition Research Council. This action was taken in conformity with the expanded objectives of the Council which will include the study of animal nutrients other than vitamins. Plans were discussed for the collaborative investigation of the toe-ash procedure in the A.O.A.C. chick assay for vitamin D. This procedure, if officially adopted, will substantially reduce the time necessary for the completion of vitamin D assays which now entail the use of solvent-extracted dried tibiae. Conferences were held on various other research projects relating to the biological assay of animal nutrients to be carried out under the sponsorship of the Animal Nutrition Research Council. Officers elected for the new term are: Chairman, KENNETH MORGAREIDGE, Director of Research and Control Laboratories, Vitamin Division, National Oil Products Co., Harrison N. J.; Secretary, FULLER D. BAIRD, Standard Brands, New York City; Treasurer, GEORGE H. KENNEDY, E. I. du Pont de Nemours, New Brunswick. In addition to the above, new members of the Executive Committee include C. I. BLISS, HERBERT C. SCHAEFER, R. V. BOUCHER, and H. R. HALLORAN. . . .

A NOTE FROM THE EDITORIAL COMMITTEE

This issue of the Biometrics Bulletin completes the second year of publication. There are 1150 members and 150 subscribers. This indicates that an active interest exists in statistical methods for biological research workers.

Your opinions have been difficult to secure, because the letters to the editor have been too few. Those letters containing criticisms which were received have been helpful. On the basis of these criticisms, the Committee decided to change the format of the Bulletin. Several changes will be noted in this issue. The articles are printed in one instead of two columns, thus enabling the tables and equations to be presented more satisfactorily. The type size has been increased along with wider margins. An effort will be made to provide a better appearance and easier reading. The self cover will be continued for the time being. If the Bulletin is expanded into a journal, a few other format changes might be advisable.

The Editorial Committee has decided to shift to a quarterly publication beginning with Vol. 3. The total number of pages currently used for six issues will be maintained or increased. Voluntary services of the Committee have been taxed heavily with publication deadlines every other month.

Consideration is being given to the feasibility of establishing a Biometrics Journal. When the backlog and flow of articles is sufficient to maintain a journal, this question will be given consideration. Meanwhile expressions on the need for a journal on statistical methodology for research workers in biology would be helpful. What do you want?

Please continue to send queries to Professor G. W. Snedecor and news items to the Chairman of the Editorial Committee.

We wish to take this means of expressing our appreciation for your cooperation.

GERTRUDE M. COX, *Chairman*
Editorial Committee

BIOMETRICS BULLETIN

Contents of Volume II

INDEX

STATEMENT OF THE OWNERSHIP, MANAGEMENT, CIRCULATION, ETC., REQUIRED BY THE ACTS OF CONGRESS OF AUGUST 24, 1912, AND MARCH 3, 1933, OF Biometrics Bulletin, published bimonthly at Washington, D. C., for 12 months ending October, 1946.

Washington, D. C., ss. Before me, a Notary Public in and for the State and county aforesaid, personally appeared Lester S. Kellogg, who, having been duly sworn according to law, deposes and says that he is the managing editor of the Biometrics Bulletin and that the following is, to the best of his knowledge and belief, a true statement of the ownership, management (and if a daily paper, the circulation), etc., of the aforesaid publication for the date shown in the above caption, required by the Act of August 24, 1912, as amended by the Act of March 3, 1933, embodied in section 537, Postal Laws and Regulations, printed on the reverse of this form, to wit:

1. That the names and addresses of the publisher, editor, managing editor, and business managers are: Publisher, American Statistical Association, 1603 K Street, N. W., Washington 6, D. C.; Editor, Gertrude M. Cox, Institute Of Statistics, Raleigh, N. C.; Managing Editor, Lester S. Kellogg, 1603 K Street, N. W., Washington 6, D. C.; Business Managers, None.

2. That the owner is: (If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding one per cent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a firm, company, or other unincorporated concern, its name and address, as well as those of each individual member, must be given.) American Statistical Association, 1603 K Street, N. W., Washington 6, D. C.

3. That the known bondholders, mortgagees, and other security holders owning or holding 1 per cent or more of total amount of bonds, mortgages, or other securities are: (If there are none, so state.) None.

4. That the two paragraphs next above giving the names of the owners, stockholders, and security holders, if any, contain not only the list of stockholders and security holders as they appear upon the books of the company but also, in cases where the stockholder or security holder appears upon the books of the company as trustee or in any other fiduciary relation, the name of the person or corporation for whom such trustee is acting, is given; also that the said two paragraphs contain statements embracing affiant's full knowledge and belief as to the circumstances and conditions under which stockholders and security holders who do not appear upon the books of the company as trustees, hold stock and securities in a capacity other than that of a bona fide owner; and this affiant has no reason to believe that any other person, association, or corporation has any interest direct or indirect in the said stock, bonds, or other securities than as so stated by him.

5. That the average number of copies of each issue of this publication sold or distributed, through the mails or otherwise, to paid subscribers during the twelve months preceding the date shown above is (.....). (This information is required from daily publications only.)

LESTER S. KELLOGG, *Mg. Ed.*

Sworn to and subscribed before me this 8th day of October, 1946.
(My commission expires July 31, 1949.)

BEVERLY RUTH RISTON, *Notary Public.*

Mr. Louis Weiner
125 Worth St.
New York 13, N. Y.

B-B